EXHIBIT 4

3

Statistical and Psychometric Issues

The administration, scoring, and interpretation of neuropsychological tests are major sources of information used in the clinical practice of a neuropsychologist to make decisions about patients' cognitive status, diagnosis, prognosis, and treatment. However, accurate decisions based on test results cannot be made without a clear understanding of the issues related to the measurement of psychological phenomena and the statistical properties of the tests. This chapter reviews basic statistical concepts of importance to neuropsychologists. No intent was made to provide a comprehensive review of statistics. The goal of this chapter is to help a novice understand and interpret psychometric data.

MEASUREMENT AND INTERPRETATION OF NUMERICAL VALUES

Measuring abilities and traits is an inherent part of clinical work. It facilitates decision making by relating the performance of a given individual to an appropriate reference group or by uncovering a change in an individual's behavior over time. The nature of decision making is specific to a given situation and includes a wide range of decisions, such as identifying cognitive strengths and weaknesses, choosing an appropriate course of cognitive rehabilitation, evaluating a patient's ability to function independently in an everyday environment, choosing an academic field and professional career, making diagnostic differentiation between disorders that affect cognitive functioning, assessing the rate of improvement or deterioration in functional capacities, and making prognostic predictions.

The concept of measurement implies numerical representation of certain properties. Such physical properties as dimensions, intensity, speed, and gravity, which represent a core of scientific exploration, lend themselves to accurate and reliable measurement. In contrast to direct measurement of physical phenomena, psychological attributes such as cognitive abilities, personality traits, and emotional status cannot be measured directly. To assess these psychological constructs, we need to obtain a sample of behavior that can be quantified and represented in numerical scores. Well-validated psychological tests are designed to elicit behaviors that are representative of the underlying psychological constructs.

Numerical values derived from an individual's performance on a test are identified as raw scores and may represent the number of correct responses, time required for completion of the test, number of errors, rating of the quality of a drawing, or different combinations of the above criteria.

In contrast to physical measuring scales that have an absolute 0 point, scaling of psychological measures does not start at the point of "no ability at all"—i.e., a raw score of 0 on an arithmetic test does not indicate that the patient has no ability to solve arithmetic problems. If the test included basic operations of addition or subtraction of single digits, the patient would be likely to succeed on these items. As a result, we cannot infer that a patient who received a raw score of 50 on a test is twice as good at arithmetic as someone with a score of 25. Due to the lack of the absolute 0 point in psychological measurements, ratios of scores are meaningless and most psychological tests are scored on an interval scale. Despite some disadvantages of the interval scale in comparison to the ratio scale, both of these scales provide measurements that lend themselves to advanced statistical analyses.

The raw score obtained on a test says little about an individual's level of ability or mastery of the subject. To interpret the raw score, it should be related to the content of the test or compared to the performance of a group of individuals on the same test. Various interpretive strategies are outlined in Chapter 2; however, the statistical and psychometric issues related to the most frequently used interpretation strategies are discussed below.

Test performance interpretation can be based on different reference criteria:

1. *Domain-* or *content-referenced interpretation* indicates how proficient an individual is in the domain tapped by the task presented on the test. Content mastery is usually reported as a percentage of correct responses on the test.
2. *Criterion-referenced interpretation* relates an individual's performance on a test to an external criterion measure, such as a practical situation which requires skills assessed by the test. For example, expectancy tables tie different levels of test performance to expected practical outcomes.
3. *Norm-referenced interpretation* compares scores achieved by one individual to the performance of a respective group of individuals who have similar characteristics. This normative or standardization sample is assumed to be representative of the population from which it is drawn and is used as an external standard of performance for interpretation of individual scores. There are several methods of relating individual performance to the norms:
   a. Raw scores obtained on a test can be converted to age or grade equivalents, which allow interpretation of a particular score in the context of expected performance for a specific age or grade level. This method is highly useful in assessing the developmental standing of children in comparison to their peers, provided that a considerable and well-identified increment in ability with age and grade advancement is expected. However, this method loses its effectiveness when the rate of development becomes uneven and the relationship between levels of ability and developmental markers weakens.
   b. Measures of relative standing of individual scores within a distribution provide an alternative method of evaluating individual performance. Percentile rank (PR) reflects the percentage of the standardization sample that scored lower than the individual score (plus one-half of that portion of the standardization sample who achieved the same score as the individual being assessed). The PR is useful for providing the relative standing of a score; however, it indicates only the ordinal position of the individual score within the distribution. It does not show dispersion of the remainder of the distribution below that score and does not indicate the absolute amount of difference between scores. For example, percentile transformations magnify differences between individuals close to the center of the distribution and compress the differences at the extremes.

   Let us consider a distribution of scores *(A)* representing the number of words recalled on the fifth trial of the

Rey Auditory-Verbal Learning Test (RAVLT):

$A = (5, 7, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 12, 13, 15)$

There are 21 observations in this distribution. An individual who obtained a score of 9 performed better than four individuals who received scores lower than 9, plus two out of four individuals (0.5) who achieved a score of 9. Relating this proportion of six individuals to the total number of 21 observations yields a PR of 29, as follows from the formula:

$$PR_{(9)} = \frac{4 + (0.5)\,(4)}{21} \times 100 = 29$$

Similar calculations of PR for scores 8 and 7 in the above distribution yield 14 and 5, respectively:

$$PR_{(8)} = \frac{2 + (0.5)(2)}{21} \times 100 = 14$$

and

$$PR_{(7)} = \frac{1}{21} \times 100 = 5$$

As follows from the above calculations, the difference in PR between scores 9 and 8 ($29 - 14 = 15$) is greater than the difference between scores 8 and 7 ($14 - 5 = 9$). This example illustrates the main disadvantage of PRs: whereas they reflect the position of an individual score relative to the standardization sample, they do not indicate the absolute differences between scores.

c. To accommodate the absolute differences between scores, interpretation of a raw score should be based on the relative standing of the score with respect to the mean for the distribution and the variability of the scores within the distribution. This can be accomplished through converting a raw score into a standard score. The most frequently used standard scores are z and T scores.

STANDARDIZATION OF RAW SCORES

Compare the distribution of RAVLT scores *(A)* used in the above example with another distribution of scores on this test *(B)*, both of which range from 5 to 15 with a mean of 10:

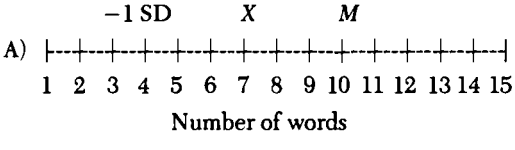
$A = (5, 7, 8, 8, 9, 9, 9, 9, 10, 10, 10, 10, 10, 11, 11, 11, 11, 12, 12, 13, 15)$

$B = (5, 5, 5, 5, 6, 6, 6, 7, 7, 8, 10, 12, 13, 13, 14, 14, 14, 15, 15, 15, 15)$

Visual examination of these distributions suggests that the variability around the mean is much greater for distribution *B*. Therefore, a score of 7 would indicate very poor performance relative to distribution *A* and a much better performance relative to distribution *B*. To account for the degree of variability in the normative distribution, individual measurements are converted into z scores.

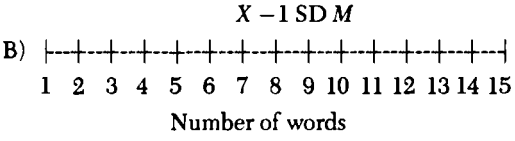
In another example, assessing recall of an individual on the fifth trial of the RAVLT, we use a reference sample with a mean of 10 (see graphs below). If the individual recalled seven words, comparison of the raw score with the mean for the reference sample ($X - M$) suggests that this individual's recall was three words below the expected score of 10 for his or her age. However, this does not tell us how low his or her performance was relative to the distribution of the normative sample.

If a high degree of variability is expected and the standard deviation (SD) for the reference sample is 6 (graph A), then a score of 7 falls halfway between a mean of 10 for the reference sample and a score of 4 representing 1 SD below the mean, which results in a z score of -0.5.



In reference to another sample with the same mean but a much lower degree of variability reflected in an SD of 1 (graph B), a score of 7 lies 3 SDs below the mean ($z = -3$). Therefore, the recall of seven words indicates

much poorer performance in relation to distribution *B* than in relation to distribution *A*.



Thus, to account for the variability within the normative distribution, raw scores are standardized, i.e., converted into *z* scores that relate the difference between an individual score and the group mean $(X - M)$ to the SD for the reference group:

$$z = \frac{X - M}{\text{SD}}$$

A negative *z* score indicates that the raw score lies below the mean for the reference group, a positive *z* score represents higher performance than the mean for the group, and a *z* score of 0 indicates that the raw score is equal to the mean of the reference group.[1]

The *z* score (SD units) shows not only how much an individual performance deviates from the mean of the sample but also how likely it is that other individuals in the sample would achieve scores as high or as low as the person being tested.

Standardization of raw scores, e.g., their conversion into *z* scores, allows comparison of the relative standing of individuals across different tests in spite of the differences in the measurement scales or the means and SDs for these tests. A standardized distribution of *z* scores has a mean of 0 and an SD of 1 because the mean is subtracted from each score and the result is divided by the SD. It preserves the same shape as the distribution of the raw scores from which it was derived. Therefore, differences in standard scores are proportional to the differences in the corresponding raw scores.

[1]For those tests that measure performance in terms of time or number of errors, where the higher scores reflect lower performance, z scores represent an inverse of the obtained score. Mathematically, in these cases the numerator should be multiplied by −1, i.e.−(X − M).

In spite of the obvious advantages of using *z* scores over raw scores, some of the properties of *z* scores are viewed as undesirable: *(1)* *z* scores have fractional values, which are carried to at least one decimal place; *(2)* half of the *z* scores in the standardized distribution are negative and half are positive, which leads to the zero-sum problem (i.e., corresponding values on both sides of the distribution cancel each other when totaled).

Parameter values of the standard distribution are arbitrarily designated. Therefore, they can be easily changed through simple arithmetic transformations of *z* scores. *T*-score transformations overcome these disadvantages through multiplying *z* scores by 10 (thus eliminating fractional values) and adding a constant of 50 (which eliminates negative values and places all the scores on a scale of 0–100 with a mean of 50 and SD of 10):

$$T = 10z + 50$$

For example, a *z* score of −1.6 can be expressed in *T* scores as follows:

$$T = (-16) + 50 = 34$$

An example of a test which uses *T*-score conversion is the Minnesota Multiphasic Personality Inventory (MMPI) and its recent revision. Clinically significant elevations on the scales are judged relative to a mean of 50 and an SD of 10, which equates the scale of measurement across all validity and clinical scales on this test.

STANDARD SCORES AND NORMAL DISTRIBUTION

Many biological measures and human characteristics are distributed so that the highest frequency of scores is observed around the distribution mean, with a gradual decrease in the frequency further away from the mean, which eventually tails off on both sides. Score distributions of many psychological tests approximate this model, which in its ideal hypothetical form represents a normal distribution. It

is convenient to treat test score distributions as if they were normally distributed because the properties of this model are known:

1. The distribution of hypothetical score frequencies arranged from the lowest to the highest values is bell-shaped and symmetrical; i.e., the left and right sides are mirror images of each other.
2. The frequency is highest in the middle of the distribution; therefore, the mean, median, and mode have the same values and divide the distribution into two equal parts.
3. The normal distribution stretches from minus to plus infinity; thus, the "tails" of the distribution get closer and closer to the x axis as they get farther away from the mean but never touch the x axis.
4. The normal distribution is described by a specific mathematical formula.

Although the test score distributions do not perfectly match this model, if the number of cases were increased and smaller class intervals were used, the shape of the sample distribution would become relatively smooth and symmetrical, thus approximating the distribution of scores in the population from which the sample was drawn.

This assumption of normality of the test score distribution allows it to be converted into a distribution of z scores with a mean of 0 and an SD of 1, which represents a standard normal distribution. Use of this conversion facilitates interpretation of the test scores because it allows comparison of a variety of otherwise not comparable distributions through equating their means and SDs. The proportion of cases comprising a certain area under the curve between two points along the z axis is known, which permits conversion of z-score units into percentiles. For example, it is known that 34.13% of all scores lie between $z = 0$ and $z = +1$. Since the mean of the distribution ($z = 0$) divides the distribution in half, we know that 50% of all scores lie below the mean. Thus, adding 34.13% of scores above the mean to 50% of scores below the mean suggests that the 84.13th percentile corresponds to $z = +1$.

Figure 3.1 illustrates the corresponding conversion values for selected z scores. The proportion of scores (i.e., the area under the standard normal curve) for each value along the z axis can be easily determined using tables provided in any basic statistics textbook.

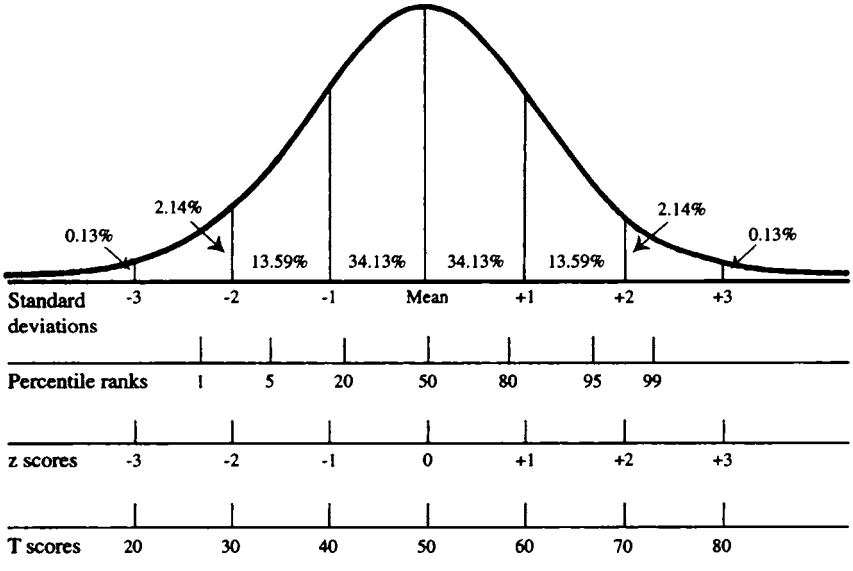


Figure 3.1. Illustration of the relationship between proportion of scores (represented by the area under the normal curve), percentile ranks, z scores, and T scores.

INTERPRETATION OF INFREQUENT (OUTLYING) SCORES

As follows from Figure 3.1, 68.26% of all scores fall within 1 SD from the mean in both directions, 95.44% fall within 2 SDs, and almost all scores except for 0.0026% are included between −3 and +3 SDs from the mean.

This correspondence between the proportion of cases and *z*-score values is important for interpretation of individual test performance since such interpretation is based on the relative frequency of the score obtained by the individual being assessed with respect to the distribution of scores. For example, a test score falling outside the range of 2 SDs above or below the distribution mean is highly infrequent; only 4.56% of the scores deviate that far from the mean in both directions. Therefore, individuals obtaining these infrequent scores may be viewed as outliers. The decision criteria for defining scores as outlying might vary from more conservative to more liberal in different clinical situations, depending on the cost–benefit ratio of making false-positive vs. false-negative errors.

Outlying scores can have different origins:

1. They might be due to the inherent variability in the population. Indeed, in any population, innate levels of a trait or ability range from very low to very high, which is modeled by a normal distribution. Therefore, a certain proportion of extreme scores is a natural feature of the population.

2. There might be purely deterministic reasons accounting for too low or too high scores, such as the following:

a. Inadequate reliability of the measuring instrument
b. Variations in test administration and scoring strategy
c. Errors in data recording or in calculating appropriate statistics
d. Demographic factors and physical handicaps affecting performance
e. Situational factors (e.g., external noise)
f. Sensitization and anxiety associated with the testing situation
g. Emotional factors and level of endurance (fatigue)
h. Response style and response bias
i. Motivational factors
j. Previous exposure to similar tests (practice effect)

3. Outlying scores may result from an execution error. In this case, an individual score is foreign to the distribution used for comparison. For example, an elderly individual of low average ability might appear impaired when compared to a normative sample of highly functioning, independently living, relatively healthy elderly individuals. This bias in the normative sample, which is not representative of all functional and economic levels of the population for the respective age group, would result in an inflated mean and in upward "slippage" of the entire distribution. To avoid execution errors, a clinician should be highly sensitive to the appropriateness of the norms used for each individual being evaluated.

INTERPRETATION OF SCORES THAT ARE NOT NORMALLY DISTRIBUTED

Interpretation of individual test scores with respect to the normative distribution is based on an assumption of normality of this distribution. To avoid interpretive errors, the basis for test score interpretation should be different if distribution is asymmetrical. Standardized distribution has the same shape as the original distribution of test scores, which is highly dependent on the characteristics of the individual items comprising the test.

If a test is designed in such a way that the majority of individuals can succeed on most of the items, test scores are compressed into a few discrete values at the upper extreme of the score range, with only a few observations at the lower part of the score range. In this case, the distribution is negatively skewed and the variability of scores falling within or above the normal range is highly limited. The test has its highest discriminative power at the lower ability levels; i.e., it is most useful in identifying impaired individuals. For example, the

distribution of scores on the Boston Naming Test and the Rey-Osterrieth Complex Figure (copy condition) in a sample of highly functioning individuals would acquire this shape.

When test items present difficulty for the majority of subjects, the score distribution is positively skewed. Variability within the lower range of scores is highly limited, whereas the highest sensitivity is obtained in the upper part of the distribution. Such a test would be most appropriate for the selection of a few outstanding individuals from a large group. The distribution of scores for Raven's Advanced Progressive Matrices can be used as an example.

In both of the above cases resulting in skewed score distributions, use of *z*-score conversions is inappropriate since such conversions are based on the assumption of normality (particularly symmetry) of the distribution.

PSYCHOMETRIC PROPERTIES OF TESTS

In view of the proliferation of psychological tests with a high overlap in terms of abilities assessed, we are frequently faced with a dilemma: which test to select in a particular situation. All published tests have to meet the requirements outlined by the *Standards for Educational and Psychological Testing* (American Psychological Association, 1999). Yet, the choice of a test should be made in the context of a certain clinical situation. In choosing an appropriate test, one has to keep in mind several criteria for evaluating its psychometric properties.

Reliability

When measuring a certain aspect of functioning, our main assumption is that the scores on a particular test would be consistent over repeated administrations. If an individual sometimes receives high scores and sometimes low scores on the same test, no inferences can be made regarding the level of ability being measured. In other words, we have to be assured that the test is a reliable measure of a stable construct such as a specific ability. However, a certain degree of variability is inherent in test performance. It is due to transient factors associated with the testing situation and the patient's state at the time of testing.

Thus, the score on a test reflects the contribution of the following two factors:

$$X = T + e$$

where X is the score on a test, T is a "true" score representing the actual level of ability measured by a test, and e is an error of measurement reflecting random variability.

With an increase in test reliability, a considerable proportion of the variability in test scores is due to differences between subjects in the "true" scores. In other words, reliability can be expressed as the proportion of variance in test scores which is accounted for by the "true" differences between subjects on the ability being measured. Therefore, a reliability coefficient provides a measure of test reliability representing the ratio of "true" score variance to the total variance of the test scores:

$$r_{xx} = \frac{\sigma^2 T}{\sigma^2 T + \sigma^2 e}$$

Methods of Estimating Test Reliability

Several procedures have been developed to determine the reliability of a test by measuring the proportion of "true" variance vs. the proportion of "error" variance. Different methods define measurement error with respect to different sources of error. The four most common methods are described below:

1. The *test–retest method* assesses the consistency of test scores from one test administration to the next. It is measured as the correlation between the scores on the first test and the retest and reflects the stability of scores over time.
2. The *alternate forms method* assesses the correlation between scores obtained by the same subject on alternate forms of a test. This method is the closest approximation to the parallel tests model.

3. The *split-half method* involves splitting the test into two equivalent halves after a single administration. There are different ways of splitting a test. The highest comparability of the two halves is achieved by an odd–even split in which one form contains all odd-numbered items and the other form, all even-numbered items.
4. The *internal consistency method* estimates the reliability of a test based on the number of items and the averaged intercorrelations among them. This method is mathematically related to the split-half method. Coefficient α is the most general form of this method and represents the mean reliability coefficient obtained from all possible split-half comparisons. In essence, internal consistency estimates compare each item on a test to every other item.

There is no universally agreed best method to evaluate test reliability. Each method has its advantages and disadvantages. The split-half reliability method overcomes theoretical and practical problems associated with the test–retest and alternate forms methods, such as difficulty in developing two equivalent forms of a test, carry-over effects, reactivity effects, and the effect of random variability on two test probes. However, the reliability estimate obtained by the split-half method varies depending on the arbitrarily chosen method of splitting. In addition, the split-half reliability coefficient underestimates the reliability of the full test and requires the use of a correction formula.

The level of reliability varies for different tests. Ideally, a highly reliable test would be preferred to a test with low reliability. However, many practical considerations might influence a clinician's test selection. The cost of error in a decision-making situation is another factor which needs to be considered in selecting an appropriate test for the given situation. Test reliability should be high when a patient's test performance is considered as one of the factors in making a final diagnostic determination. Tests with lower reliability might be acceptable in preliminary screening situations.

Typical levels of reliability attained by neuropsychological tests range from 0.95 to 0.80, which represents a high to moderate range of reliability. For a test with a reliability estimate of 0.80, 20% of the variability in scores is due to measurement error. Thus, tests with reliability below 0.80 introduce a considerable proportion of "noise" in scores, which compromises their interpretability. For screening tests, reliability between 0.80 and 0.60 would be acceptable, whereas reliability estimates below 0.60 are usually judged as unacceptably low.

Standard Error of Measurement

The reliability estimate provides a relative measure of the accuracy of test scores. As any correlation, it is influenced by the variability of scores. In a sample with a heterogeneous score distribution, reliability will be higher than in a more homogeneous sample.

The reliability estimate does not indicate how much variability should be expected due to measurement error and how accurate the individual test scores are. Therefore, in addition to reporting reliability coefficients, test developers report the size of the standard error of measurement (SEM), which is useful in interpreting the observed scores of each individual patient. The SEM is determined by the reliability of the test (r_{xx}) and the variability of test scores (σ_x):

$$\text{SEM} = \sigma_x \sqrt{1 - r_{xx}}$$

Since no test provides a perfect measure of ability, a certain degree of variability in the scores obtained by the same subject is expected. The SEM indicates how much an individual's score might vary if he or she is retested repeatedly with the same test (assuming that there is no practice effect or fatigue effect). According to measurement theory, the scores obtained by one subject across an infinite number of retests with the same test would result in a normal distribution, with the mean equal to this subject's "true" score and the SD equal to the SEM.

Since in most clinical situations we obtain only one score on a test, we may treat it as an estimate of the theoretical "true" score. Using

the SEM, we can form a confidence interval (CI) around this score, which provides a range in which a subject's "true" score is likely to fall.

For example, a 70-year-old patient obtained a WAIS-III Full Scale IQ (FSIQ) of 110. According to the test manual, the size of the SEM for FSIQ in this age group is 2.19 (Wechsler, 1997). Since we know that 95% of all scores in a normal distribution fall within 1.96 SD of the mean, the 95% CI for this score will fall between 1.96 SEM below and 1.96 SEM above the obtained IQ score (110 ± 1.96 SEM). Calculation of a CI by multiplying the SEM by 1.96 ($2.19 \times 1.96 = 4.29$) suggests that we would expect 95% of the IQ scores obtained by this patient to fall in the range of 110 ± 4.29, or between 105.71 and 114.29.

If we want to increase the level of certainty in constructing a range in which a patient's true score is likely to fall, we can use the 99% CI. In other cases, we might use lax CIs that provide accuracy below the 95% level.

The drawback in using SEMs to determine the accuracy of test scores is the fact that they do not have the same size for all scores: they are smaller for extreme scores and larger for moderate scores. Another limitation in using SEMs is that test scores are generally further away from the mean than "true" scores because of the tendency for regression toward the mean. To overcome this distortion, the CI can be formed around the estimated "true" score, which is obtained from a regression equation.

Validity

When a test is used to assess a certain aspect of functioning, it is assumed that the test measures what it is supposed to measure and that it is useful in making accurate decisions. Different validation strategies are used to understand the meaning and implications of the scores achieved on the test. Content validity and construct validity indicate whether a test is a valid measure of a specific ability. Criterion-related validity refers to the accuracy of decisions that are based on the test scores.

1. *Content validity* reflects the extent to which the behaviors sampled by the test are a representative sample of the ability being measured. It is not measured statistically but determined by agreement among expert judges with respect to a detailed description of the content domain that is measured by each test item.
2. *Construct validity* determines how well observable behaviors measured by the test represent underlying theoretical construct. This relationship can be established through high correlation of the test with other tests measuring the same construct *(convergent validity)* or through low correlation with tests measuring different constructs *(discriminant validity)*. If more than one method is used to measure several constructs, the correlations among them can be represented in a multitrait–multimethod matrix, which establishes whether the results of a certain test are determined by the construct being measured or by the method of measurement. Construct validity can be further assessed by correlating one test with many other tests using factor analysis. In this case, construct validity is established through the high loading of a particular test on the factors that represent those constructs presumed to be measured by the test.

   Thus, content validity and construct validity represent two different strategies in determining that the test measures what it is supposed to measure: "Content validity is established if a test *looks* like a valid measure; construct validity is established if a test *acts* like a valid measure" (Murphy & Davidshofer, 1991).
3. The usefulness of a test in decision-making situations represents another aspect of test validity. *Criterion-related validity* reflects the relationship between test scores and measures of decision outcome, i.e., criteria. Any measurable behavior can be used as a criterion. For example, the choice of a rehabilitation strategy can be evaluated using a measure of symptom reduction as a criterion, or the accuracy of a screening test can be

assessed using a patient's psychiatric diagnosis as a criterion. The correlation between test scores and the criterion measure, which is derived without using the test, reflects the accuracy of predictions or decisions made on the basis of the test scores.

Criterion measures can be obtained *after* decisions are made based on the test scores in a random sample of a population about which decisions are made *(predictive validity)* or *at the same time* that decisions are made in a preselected sample *(concurrent validity)*. Whereas predictive validity is superior to concurrent validity in that it is a direct measure of the relationship between test scores and a criterion measure for the general population, it has a number of practical and ethical drawbacks. For this reason, the most practical and commonly used measure of criterion-related validity is concurrent validity, despite the fact that its coefficient underestimates the predictive validity.

Theoretically, an estimate of the correlation between test scores and a criterion measure obtained in a criterion-related validity study can range between −1 and +1. Validity coefficients for most of the tests are relatively low, ranging between 0.2 and 0.5. This is due to the imperfect reliability of the test and the criterion measure: whereas a criterion is assumed to represent the "true state" of a patient, it is frequently based on subjective clinical judgment, which is inherently unreliable.

If the correlation coefficient between a test and a criterion is 0.3, the proportion of the variance in the criterion that is accounted for by the test (r^2, or coefficient of determination) is 0.09. This means that only 9% of the variability in the criterion can be accounted for by the test scores. Although these numbers look discouraging, they should be interpreted in the context of other measures that contribute to the accuracy of decisions.

Decision Theory

In clinical practice, a clinician has to make decisions which range from assigning a certain diagnosis to applying a specific course of treatment. Since predictions based on the information available to the clinician are never perfect, each decision may have several possible outcomes. In the context of decision theory, the predictor and criterion values are reduced to only two categories, in spite of the continuous nature of these values. Comparison of predictions with criterion values suggests that there are four possible outcomes of decisions: correct decisions include true-positive (TP) and true-negative (TN) outcomes, whereas incorrect decisions include false-positive (FP) and false-negative (FN) outcomes. Tests are used to maximize the number of correct decisions and to minimize the number of errors. The contribution of the criterion-related validity of a test to improvement in the accuracy of decisions depends on the base rate and selection ratio.

Base rates

The base rate reflects the proportion of an unselected population who meet the criterion standard. Clinically, this term is used interchangeably with *incidence* or *prevalence* of a disorder. In a hypothetical example, assume that among 500 normal elderly, 9% would have scores below the cutoff for dementia of 24 on the Mini-Mental State Exam (MMSE). Assume that 80% of patients with the diagnosis of dementia of Alzheimer's type (DAT) score below a cutoff of 24. If 100 DAT patients were added to 500 intact elderly (total number of subjects 100 + 500 = 600), the base rate would be 100/600, or 17%. In this situation, the outcomes would be as follows:

$$100 \times 80\% = 80 \text{ TP}$$

$$500 \times 9\% = 45 \text{ FP}$$

$$500 - 45 = 455 \text{ TN}$$

$$100 - 80 = 20 \text{ FN}$$

Thus, follow-up of subjects who score below the cutoff (80 TP + 45 FP = 125) will yield

a hit rate of 80/125 = 64%. In other words, the diagnosis of DAT will be confirmed in 64% of those subjects who scored below the cutoff of 24 on this test.

In contrast to the above hypothetical example, in the general population, the base rate for DAT is considerably lower than 17%. For example, if the base rate for DAT is 5%, then out of 600 subjects, 30 would be suffering from DAT and 570 would be intact with respect to this diagnosis. Assuming that 80% of DAT patients and 9% of the intact elderly score below the cutoff on the MMSE, as was the case in the above example, the table of outcomes would look different because of the lower base rate for DAT:

$$30 \times 80\% = 24 \text{ TP}$$

$$570 \times 9\% = 51 \text{ FP}$$

$$570 - 51 = 519 \text{ TN}$$

$$30 - 24 = 6 \text{ FN}$$

The ratio of TP scores to the total number of subjects scoring below the cutoff (24 TP + 51 FP = 75) will yield a hit rate of 24/75 = 32%, which is considerably lower than in the example with a higher base rate.

With a decrease in base rates, most of the population are negatives; positives become more rare, and therefore, an attempt to identify this group will lead to an increase in the number of FP decisions. Low base rates also lead to a large number of TN decisions since a majority of the population do not suffer from DAT. Following the same logic, in the case of high base rates, as the number of TP decisions increases, the frequency of FN errors also increases. An optimal base rate of about 50% minimizes decision errors and maximizes accurate decisions, providing that the test used to assist in decision making has sufficient validity.

In the general population, the base rates for certain disorders are usually low and most of the "red flags" represent false alarms. Base rates among individuals referred for evaluation due to progressive symptomatology are higher, and therefore, the expected number of false alarms would be lower.

Selection ratio

Another factor affecting the accuracy of decisions is the *selection ratio*, which is defined as the ratio of TP + FP outcomes to the total number of subjects. Assume that a psychiatric ward has 30 beds for severely depressed patients. If only 32 patients are referred for hospitalization at any one time, the selection ratio would be high (30/32 = 0.94). The hospital cannot be very selective in this situation, and most of the referred patients could be hospitalized. In another scenario, 100 patients could be referred for hospitalization at any one time due to severe depression. Since the selection ratio is low (30/100 = 0.30), a certain strategy needs to be used to identify those who are acutely suicidal for immediate hospitalization. When a selection ratio is low, a test with even modest validity can make a considerable contribution to the accuracy of decisions.

Incremental validity

The utility of a test has to be assessed in terms of an increase in the accuracy of decisions obtained using the test, which extends beyond the base rate or beyond information obtained from other sources. In other words, incremental validity reflects the unique contribution of a test to understanding the patient.

Incremental validity is affected by the base rate, selection ratio, and criterion-related validity of the test. When decisions are made at random, the frequency of different outcomes can be computed directly from the base rate and the selection ratio. The incremental validity of a test indicates the degree of improvement in the accuracy of decisions, i.e., frequency of TP and TN outcomes, beyond the random level, which are made using the test.

The incremental validity is highest when the base rate is moderate, selection ratio is low, and criterion-related validity is high. Values of incremental validity for different combinations of base rates, selection ratios, and criterion-related validity coefficients are provided in Taylor-Russell (1939) tables.

Thus, the validity coefficient alone does not determine the usefulness of a test in each

clinical situation. Test usefulness depends largely on the context in which the test is used.

Cutoffs and diagnostic accuracy of a test or interpretive strategy

As pointed out above, in the framework of a decision theory approach, both the predictor (test) and criterion values are reduced to only two outcomes. Thus, the continuous nature of test scores is reduced to categories of pass/fail, impaired/unimpaired, etc. Selection of a cutoff point dividing a sequence of test scores into these two categories is another factor that affects the accuracy of decisions. Through manipulating the cutoff, the frequency of a certain type of correct decision can be maximized at the expense of increasing the frequency of another type of error.

For example, test sensitivity, or the ability to correctly identify impaired individuals (expressed as the ratio of TP to all impaired individuals [TP + FN]), can be increased by fixing the cutoff at a small number of incorrect responses. This will reduce the frequency of FN errors but increase the proportion of FP errors. In other words, this will assure correct identification of the majority of individuals with even mild impairment and very few misidentifications of impaired individuals as being intact. At the same time, this will yield a large number of intact individuals who will be misidentified as impaired. The costs of such misidentification include inappropriate treatment, psychological distress, and adverse social/economic consequences.

On the other hand, test specificity, or the ability to correctly identify the absence of impairment (expressed as the ratio of TN to all intact individuals [TN + FP]), can be increased by setting the cutoff at a large number of incorrect responses. This will reduce the proportion of FP errors but result in a large number of FN errors. In other words, only those patients who have pronounced impairment will be identified as impaired, and very few intact individuals will be misidentified. However, many individuals with mild symptomatology will be missed. This will preclude timely therapeutic intervention which otherwise would allow stabilization or reversal of these patients' symptomatology.

Thus, manipulation of the cutoff affects the balance between sensitivity and specificity and results in different cost–benefit ratios. Based on the empirical evidence, the cutoff is usually set at a value that ensures a reasonable balance between sensitivity and specificity so that only "borderline" patients will likely be misidentified. Setting the optimal cutoff yields the highest Hit Rate, i.e., ability of the test to correctly identify the presence and absence of impairment (expressed as the ratio of [TP + TN] to all individuals in the sample [TP + FP + FN + TN]).

In making a diagnostic decision, the clinician is concerned with the utility of a test in correctly identifying impairment in an individual patient, i.e., in the test's predictive value, rather than in its accuracy in discriminating between groups. Positive Predictive Value represents the probability that the patient is indeed impaired, given an impaired test score (expressed as the ratio of TP to all individuals identified by the test as impaired [TP + FP]). Negative Predictive Value represents the probability that the patient is intact given a non-impaired test score (expressed as the ratio of TN to all individuals identified by the test as non-impaired [FN + TN]).

The probability of the condition based on the test result (predictive value) is referred to as the *posttest probability*. However, the usefulness of a test in aiding diagnostic decisions is also determined by the base rates (prevalence) of the condition in a given setting (see above), which represents the *pretest probability*. These probabilities can be converted into odds of having the condition, which are expressed as the ratio of the probability of having the condition to (1−probability of having the condition). Posttest odds (which represent the likelihood that the individual who obtained a score X on the test has the condition) take into account the pretest odds and likelihood ratio:

$$\textit{Pretest odds} \times \textit{Likelihood ratio} = \textit{Posttest odds}$$

where the likelihood ratio represents the odds of a specific test result occurring in an individual who has a condition over the odds of that test result occurring in an individual who does not have the condition. In other words, it

represents the likelihood that, given a score X on the test, an individual is impaired vs. unimpaired. Likelihood ratios relate the specificity and sensitivity of a test to a given setting. They can be defined over the range of possible values, identifying a degree of abnormality rather than representing the presence/absence dichotomy, and therefore allow consideration of the greater predictive power of scores at the extremes of a distribution. Likelihood ratios are particularly useful in determining the probability of having a condition based on the results of a test battery, rather than an individual test score.

An alternative to tests of statistical significance between groups is the Odds Ratio statistic, which reflects the association between the incidence of a condition given specific situational factors versus incidence of that condition in the absence of those factors and approximates relative risk estimates when the incidence of the condition is low. Conversely, it measures the strength of dissociation between individuals with different test results and reflects the probability that the condition is present in an individual with an abnormal test result. In other words, it shows that an individual who obtained an abnormal score (falling below the cutoff for impairment) on a specific test is X times more likely to have the condition than an individual who scores in the nonimpaired range (above the cutoff).

For further discussion of the utility of diagnostic tests, see Fletcher et al. (1996) and Sackett et al. (2000).

SYNTHESIS OF RESULTS OF DIFFERENT STUDIES IN A META-ANALYSIS

Historical Overview and the Rationale for Using Meta-Analysis in This Book

Historically, researchers and clinicians struggled with the amount and diversity of information available in the literature on any topic. Efforts to summarize results of studies in medicine can be traced back to the 18th century, marked with the conception of two journals published in Germany that provided critical appraisals of new publications. William Withering's summary review on the use of digitalis for treatment of heart disease, published in 1785, is one of the first examples of a systematic review.

Statistical methods to combine data from different studies in medicine were first introduced in 1904 by British mathematician Karl Pearson. These early efforts spearheaded the development of statistical methods to synthesize research findings in social sciences, particularly in psychology and educational research. In 1976, the American psychologist Gene Glass coined the term "meta-analysis" to describe research synthesis based on statistical techniques. In the 1980s, meta-analysis became popular in medicine, particularly for summarizing results of clinical trials addressing the effectiveness of treatment techniques and of observational (epidemiological) studies examining the accuracy of diagnostic methods.

In response to a pressing concern regarding the lack of summary reviews for those who need to use evidence from unmanageable amounts of information to make informed decisions in medicine, British physician and epidemiologist Archie Cochrane founded the Cochrane Collaboration in the 1990s. This international network of health-care professionals promotes accessibility of systematic reviews through maintaining registers of controlled trials and preparing/updating systematic reviews, which are published in the Cochrane Library available on the internet (Antes & Oxman, 2001; Egger et al., 2002).

To achieve a consensus across disciplines on how to report the results of systematic reviews, the conference on the Quality of Reporting of Meta-Analyses (QUOROM) was held in 1999, bringing together clinical epidemiologists, clinicians, statisticians, and researchers from the United Kingdom and North America. As a result of this conference, the QUOROM statement was published, which includes a checklist and a flow diagram that identify the type and format of information to be included in systematic reviews (Moher et al., 1999). The goal of this "gold standard" is to help readers to evaluate the quality of reports and to appraise the likelihood of systematic error, i.e., bias in data reporting (Shea et al., 2001).

In the past decade, neuropsychology researchers have turned to comparative (case-control) observational meta-analytic techniques in an effort to examine the diagnostic accuracy of test batteries by comparing performance profiles of clinical and matched control groups. The outcome measures in such studies are sensitivity, specificity, likelihood ratio, effect size, and/or a summary receiver operating characteristic (ROC) curve. A few studies, primarily in medicine, have used noncomparative (descriptive) observational meta-analyses to identify mean values and the expected variability of a certain parameter (e.g., blood pressure) in non-diseased individuals. The studies described in this book were subjected to such non-comparative meta-analyses, the results of which are reported in the relevant test chapters. The outcome measure in our analyses is an expected distribution of scores in a nonclinical population, mediated by demographic variables, when appropriate.

Application of Meta-Analysis in Clinical Practice

The advantages and limitations of meta-analytic techniques are addressed in several publications (Cooper & Hedges, 1994; Egger et al., 2001, 2002; Green & Hall, 1984; Harris & Rosenthal, 1985; Hasselblad & Hedges, 1995; Hedges, 1982; Hedges & Olkin, 1985; Hunter et al., 1982; Kulik, 1983; Light & Pillemer, 1984; Rosenthal, 1983, 1984; Sterne et al., 2001; Sutton et al., 2000; Wolf, 1986) and are briefly summarized below.

Advantages

1. Karl Pearson was the first mathematician to point out that individual studies are too small to allow definitive conclusions, in view of the size of the probable error. To solve this problem, he proposed combining individual studies. Such a synthesis provides a solid basis for evidence-based clinical decisions in modern clinical research and practice, whereas conclusions drawn from several individual studies might be contradictory or the sample size of an individual study might be insufficient to detect or rule out a modest but important property of a specific parameter.
2. Meta-analysis identifies areas where consistent evidence is absent. It highlights the need for further research if conclusions drawn from individual studies are contradictory or high-quality studies in the area of interest are not available.
3. Analysis of heterogeneity in study results helps to identify subgroups that differ in an estimated parameter and draws attention to factors mediating the outcome.

Sources of Bias

1. Publication bias has been of considerable concern and was described by Rosenthal (1979) as a "file drawer problem." It points to the fact that studies yielding statistically significant findings are more likely to be published, published without delay, and published in English than studies with "negative" findings, which tend to be filed away. This causes a Type I publication bias error and results in a spurious effect of the parameter under investigation. To remediate this bias, the influence of unpublished studies and those published in languages other than English on the outcome should be taken into consideration.
2. Methodological and design quality differences between studies in terms of degree of experimenter blindness, randomization, sample size, controls for recording errors, and type of dependent variable (e.g., self-report vs. objective) represent another source of bias. Some researchers suggest that studies of higher quality (with larger samples sizes and well-controlled sampling) tend to have lower variance and effect sizes.
3. The issue of combinability of studies in a single meta-analysis is rooted in homogeneity of parameter estimates. Homogeneity across studies is assumed, based on the expectations that all studies are testing the same hypothesis and estimating the same population parameter

and that variations in study estimates are random.

However, heterogeneity of parameter estimates is a common problem. There are several distinct points of view on how to deal with heterogeneity:

a. Heterogeneity is analogous to individual differences among subjects within single studies and represents variations within the same parameter.
b. The studies should be grouped into homogeneous subsets and combined in separate meta-analytic syntheses.
c. Outliers contributing to heterogeneity should be subjected to close examination, to test for mediating effects that may contribute to the heterogeneity and to better understand the properties of the parameter of interest and suggest new hypotheses.

In addition to statistical methods directed at reduction of heterogeneity, a practical approach to this problem rests on treating meta-analyses differently from systematic reviews. It is expected that all available data will be systematically reviewed. However, it might be inappropriate to pool data from all heterogeneous studies in a meta-analysis. This is especially true for case-control epidemiological studies, where combining a set of confounded studies may result in spuriously precise but biased estimates of association. Thus, careful examination of the data to determine sources of heterogeneity is advocated in the literature.

4. Another source of bias stems from including multiple tests of a hypothesis from a single study in a single meta-analysis, which inflates sample size beyond the number of independent studies. A practical solution to this "apples and oranges" criticism is to perform separate meta-analyses for each type of outcome variable.

Careful attention to the sources of bias in meta-analytic synthesis enhances its reliability and validity. *Reliability* of the meta-analysis refers to reproducibility of results (i.e., the likelihood that independent meta-analysts replicating the analysis will locate and include the same studies and measures) and agreement among raters in the coding of study characteristics (Rosenthal, 1984; Wolf, 1986; Zakzanis, 1998). *External validity* and *internal validity* are directly related to the choice of studies to be aggregated (coding strategies, examining mediating effects, testing homogeneity of results) and methodological quality of the studies, respectively.

Guidelines for conducting meta-analyses to evaluate diagnostic tests and writing meta-analytic reviews have been published by Hasselblad and Hedges (1995), Irwig et al. (1994), and Rosenthal (1995).

SELECTION OF STUDIES AND PROCEDURES FOR META-ANALYSES PRESENTED IN THIS BOOK

Literature Search and Selection of Studies

The initial pool of studies considered for inclusion in the meta-analysis was generated through a computer-based search of the PSYCHINFO and MEDLINE databases. Names of the tests of interest and the neuropsychological functions measured by these tests were entered as key words in separate runs, with the search limited to English-language publications dating from 1998 until the present. The intent of this search was to add the most recent articles to a presumably comprehensive set of articles containing normative data included in the first edition of this book. References in the articles generated as a result of this search were reviewed to identify earlier relevant publications that might have been missed in prior searches. In addition, unpublished sets of normative data that have been sent to the authors after publication of the first edition of the book were evaluated for inclusion.

The meta-analytic tables are presented in this book only for those neuropsychological tests that have a sufficient number of

homogeneous studies that are based on the same version of the test or the same administration format. For those chapters that contain the meta-analytic tables, not all studies available in the literature were necessarily included in the database for the analyses. Those data sets that are based on clinical groups not well identified in terms of methodology or on administration of the tests by medical staff, rather than by trained examiners, were not reviewed. Among studies that were reviewed, those that do not contain test means and SDs (or data that can be converted into these statistics), do not have demographic descriptions of the sample, or are based on idiosyncratic samples (e.g., data collected in China) or nonstandard administration procedures were not included in the meta-analyses. An effort was made to identify multiple publications based on the same study and to include a data set from each study only once, to avoid overlapping data sets. Similarly, when data are presented in overlapping age groups, only nonoverlapping data points were used. Data sets based on medical patients and on patients referred for neuropsychological evaluation which yielded no neurological findings were not included. The resulting data sets include data collected primarily (but not exclusively) in the United States and Canada, and the vast majority of the participants across the studies are Caucasian.

Procedures Used in the Analyses

Data were analyzed using Stata, which is a general-purpose, command line–driven statistical package for data management and analysis. It reads data into storage memory and is programmable, allowing the user to add new commands. This package was used for our purposes because it contains a comprehensive set of user-written commands for meta-analysis, in addition to commonly used ordinary regression analysis tools.

Data in all analyses were inversely weighted on standard errors for the means since such weighting allows one to account for both sample size and the dispersion around the mean for each data point (calculated as a square root of the ratio of squared SD to the sample size). Data points that have a larger sample size and a smaller variance contribute more to the analysis. This helps to control for study quality since higher-quality studies tend to have larger weights. Stata's analytic weights were used, which represent the number of elements that gave rise to the statistic representing the data point.

Fixed effect with a cluster option was used for all regressions. A cluster option was used to identify data points that were derived from the same study, to account for a lack of independence of data points within each study. Ordinary least square regressions ("regress" command) were used, as opposed to the meta-analysis regression ("metareg" command), because "metareg" does not allow for the cluster option. We opted for the fixed effect based on an assumption that all data came from the same population. Preliminary tests with the "Meta" command for all data sets revealed that pooled estimates of the fixed and random effects were comparable (e.g., 42.42 and 42.22, respectively, for the FAS).

Tables of predicted values are based on the parameters identified in the above regressions and include 95% CIs (expected to include 95 out of 100 estimated values if the trials were replicated 100 times), calculated according to the following formula: 95% CI = value $\pm$ 1.96 $\sqrt{\text{var(value)}}$.

Data Editing

After the relevant literature was selected, mean test scores with their respective SDs, demographic variables, and study characteristics were recorded in the Stata database partitioned by age and/or education group or for the entire sample as reported by study authors. When data allowed gender comparisons in addition to the overall scores, they were also recorded in a separate file to avoid double sampling. Every entry in the database is viewed as a data point. For example, a study that provides test performance data stratified into 4 age $\times$ 2 education groups would generate eight data points.

Data were examined for consistency and for outlying scores. To aid us in this examination, we used the "meta" test, which tests data for

heterogeneity and assesses the influence of a single study in meta-analysis. It has an option of generating a graph depicting weighted means for all data points with their respective standard errors, centered around a vertical line demarcating the combined estimate of the mean (e.g., see, Fig. 3.2). Inspection of the resulting forest plot allows one to visualize the overall distribution of scores and to identify outlying data points. The degree of deviation from the estimated distribution parameters was further examined with a box plot and with an "iqr" test, which classifies outliers into mild vs. severe categories based on the analysis of an interquartile range. The presence of outliers typically resulted in a high Q value and a high estimate of between-study variance, representing heterogeneity among studies included in the data.

Outlying data points identified by visual inspection and through the above analyses were reviewed. Although variability across studies and age/education groups is expected, we strived to identify data included in error. Outlying data points that can be explained by clerical errors in published sources, deviations from standard administration procedures, or idiosyncratic samples were excluded from further analyses. The "meta" test was rerun on the remaining data to assure a decrease in the Q value and in an estimate of between-study variance in comparison to the initial analysis. Information related to the analyses for outlying scores is not included in the tables in the appendices, to avoid "information overload." Data reported in the tables describe heterogeneity only in the final data set, after data editing. It should be noted that the final Q values for all data sets are significant at the 0.000 level, indicating heterogeneity for all data. This outcome likely stems from the fact that the data come from different studies and

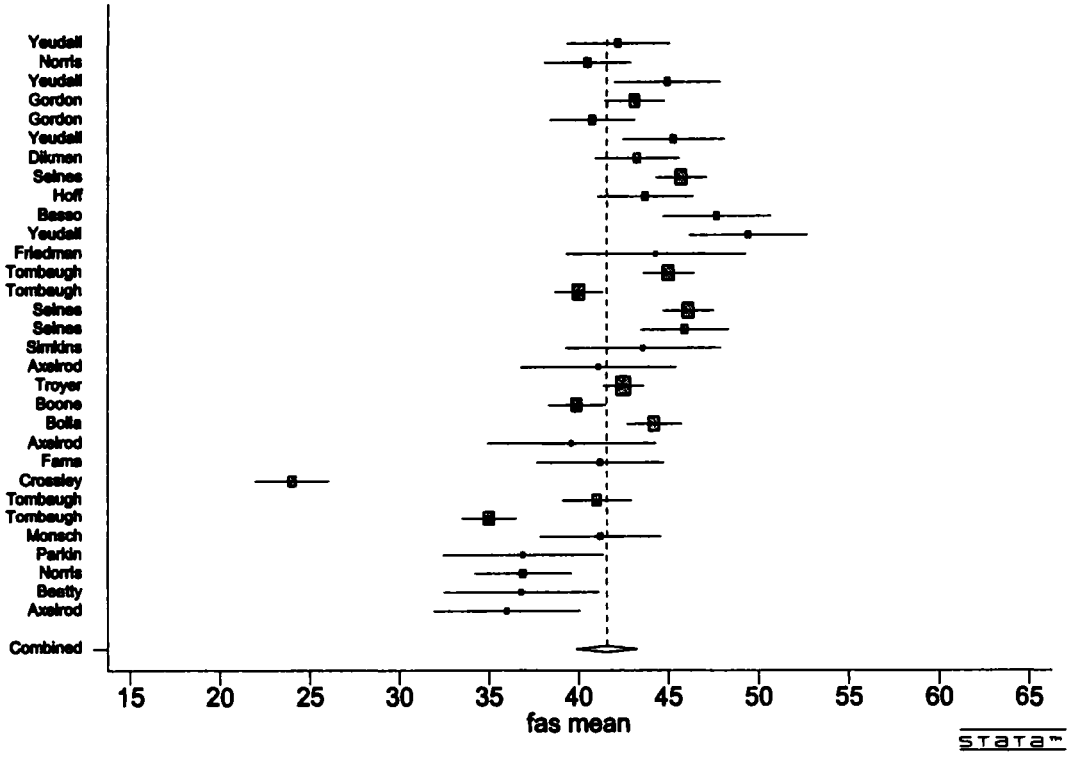


Figure 3.2. Example of a forest plot, which was used to assess the influence of a single study in the meta-analysis (data for the Verbal Fluency–FAS test were used). Vertical line demarcates the combined estimate of the mean. Data points are depicted as weighted means with their respective standard errors.

is dealt with using the cluster option (which identifies the study from which the data were derived) in the regression analyses.

Furthermore, for those tests that are sensitive to an education effect, data were examined for consistency of represented education ranges. If a large gap was detected, data falling beyond the empirically supported distribution of education ranges were excluded from the analyses, to avoid extrapolating a prediction rule over ranges that are not supported by existing data. This process is described in the respective chapters.

Regression

It has been widely documented in the literature that expected test performance varies as a function of demographic characteristics of an individual. Age has been shown to contribute most to this variability. To identify the rule describing a relationship between age and test performance, as reflected in the corresponding study means, data were subjected to regression analyses. Ordinary least square regressions with fixed effect and a cluster option were used. The shape of the distribution of means was visually inspected to assist in the decision on whether linear or quadratic regression was appropriate for a specific data set. In addition, both linear and quadratic solutions were subjected to the test for model fit. An increase in the R^2 for the quadratic model and comparison of the Bayesian Information Criterion generated for each model were used to guide the decision-making process. This information is presented in the relevant tables.

The results of the regression analysis yield rich information on the relationship between age as a predictor variable and the test means as an outcome variable. R^2 indicates the proportion of variance in the test scores accounted for by the model. It should be noted that we used R^2 rather than adjusted R^2 (which corrects for chance variation) since we had only one predictor for a relatively large number of observations in each model and, therefore, both values are very close. The F value and a corresponding probability level indicate how reliably predictor variables predict test scores. The strength of the relationship between age and the test means is reflected in the dispersion of data points around the regression line. A scatterplot illustrating the dispersion is included in each relevant chapter, with the size of the bubbles reflecting the weight of the data point (larger bubbles indicate larger standard error [SE] and smaller weight).

Information for each term of the regression model is provided in the tables. The coefficient for a predictor variable indicates the extent of gain or loss in the test performance given a one-unit change in the value of that predictor variable (given that all other variables in the model are held constant). For example, for the FAS version of the Verbal Fluency test, the coefficient for the Education term is 0.498 (see Table 11m.1 in Appendix 11m, under "Effect of demographic variables"). This means that for each 1-year increment in education we expect a 0.498-unit increment in word production. In other words, with every additional year of schooling, an individual is expected to increase verbal fluency by almost 0.5 of a word, irrespective of age. The 95% CI for the coefficient shows how high and how low the actual population value of the coefficient might be.

Dividing the coefficient by the SE for that parameter yields the t value, which is used in testing the null hypothesis that the coefficient for a given term is 0. In our example, a t value of 2.47 with a two-tailed $p = 0.025$ indicates that the coefficient for Education of 0.498 in a model based on 29 observations is significantly different from 0; thus, we can infer that education significantly contributes to test performance. This information was used in the tables to provide a correction factor for predicted test scores for different levels of education.

It should be noted that significance tests for the term *age* in the quadratic equations do not accurately reflect the linear effect of age on test performance due to collinearity with the quadratic effect, i.e., with the age^2 term in the equation. To address the linear effect of age, avoiding the collinearity, we present significance tests for *age* centered (by subtracting the mean age for the aggregate sample from

the mean age for individual samples) in the footnotes to the relevant summaries of the regression models.

Prediction

The model that was estimated using regression command was used to make out-of-sample predictions on another data set, which included values for age distributed in 5-year increments (with smaller intervals at the extremes of the age distribution in some cases), representing mathematical centers of respective age categories. For example, for the FAS, the data set includes values 19.0, 22.5, 27.5, 32.5, 37.5, 42.5, 47.5, 52.5, 57.5, 62.5, 67.5, and 72.5. These numbers represent the age categories 18–19, 20–24, 25–29, 30–34, 35–39, 40–44, 45–49, 50–54, 55–59, 60–64, 65–69, and 70–74. Care was taken to avoid out-of-range estimates. For example, when the available data extended only to the age of 82, the 80–84 category was not used in the prediction table because an assumption that the same rule applies to ages 83 and 84 would not have an empirical basis. In some cases, when a partial age group was well represented, a predicted value for this age group is listed. It should be noted that distributions of the data used for model estimation were examined for continuity, to avoid gaps within the distribution. When such gaps were detected, the extreme data points were excluded from analyses. Tables of predicted values with corresponding CIs for the relevant neuropsychological tests are presented in the meta-analytic tables in the appendices along with supporting statistics. Critical reviews of strength and limitations of predicted values are included in the text of the respective chapters.

In clinical practice, situations might arise where an estimated score is needed for an age that falls beyond the range of age categories included in a prediction table. We strongly recommend that the clinician seek the needed data in individual studies included in this book (using locator tables to facilitate the search) or turn to the data accumulated in that specific clinic. However, if everything else fails, the needed score can be calculated using the regression equations included in the tables, which underlie calculations of the predicted scores. The equations are based on the coefficients for all predictor variables used by the program. The equation for a linear model is as follows:

$$\textit{Predicted test score} = \text{constant} + (\beta_{age}) \times age$$

That for a quadratic model is as follows:

$$\textit{Predicted test score} = \text{constant} + (\beta_{age}) \times age + (\beta_{age^2}) \times age^2$$

where β_{age} is the coefficient for age and β_{age^2} is the coefficient for age^2, respectively. For example, a quadratic equation derived from the model estimation for the FAS is $34.298 + 0.554 \times age - 0.007 \times age^2$ (see Table 11m.1 in Appendix 11m). The equations are provided below the prediction tables; coefficients used in these equations are listed among the results of the analyses provided in the tables, specifically under the subtitle "Ordinary least square regression."

As reflected in the shape of the regression line in the scatterplot in Appendix 11m, FAS performance is expected to increase somewhat up to approximately age 40, with a subsequent decline. The value for age when performance reaches its maximum can be derived from the regression equation using the following formula:

$$y = -[\beta_{age}/2 \times (\beta_{age^2})]$$

Using coefficients from the regression equation for the FAS this value is $-[.554/2 \times (-.007)] = 39.57$. The obtained value represents the age at which the curve turns over to the declining direction.

Standard Deviations

To test for a possible relationship between the variability in scores at different ages, regressions of SDs for test scores on age were run. When age accounts for a significant amount of variability in the SD, the predicted values for SDs and CIs (calculated using the same approach as above) are reported along with the predicted test scores. The results of

significance tests for regressions on SDs are reported. Tests for model fit for the solutions on SDs were performed using the same approach as for the performance scores. The results of these tests were used for decision-making purposes, but they are not presented in the meta-analytic tables in the appendices, to avoid information overload.

When the results suggested that age does not account for any notable amount of variability in SD, as reflected in a very low R^2, mean SDs derived from the original data are listed in the tables as they are applicable for all age groups.

Testing Model Fit and Parameter Specifications

Postestimation tests of parameter specifications were performed to ensure accuracy of the prediction. Though violation of the normality of the residuals would not affect estimates of regression coefficients and predicted values, it would affect the validity of hypothesis testing; in other words, significant deviation from normality would affect the validity of p values for the t-test and F-test. The Shapiro-Wilk W test was used to assess the normality of residuals for the variables used in the regressions. The p value for the W statistic is based on the assumption that the distribution is normal. Thus, high values of p indicate that we cannot reject the hypothesis that the variable is normally distributed. The normality of residuals was also assessed using the "kdensity" plot (Kernel Density Estimate), which approximates the probability density of a variable, and through visual inspection of residuals regressed on age. Close approximation of the estimated curve to the normal density overlaid on the plot and no pattern in the dispersion of residuals support the results of the Shapiro-Wilk test. Kernel Density Estimate and plot of the residuals regressed on age for the FAS are reproduced in Figures 3.3 and 3.4 for illustration purposes (the size of the bubbles in Fig. 3.4 reflects the size of the SEs of the data points, reciprocal to their weights). However, they are not included in the meta-analytic tables in the appendices.

Homoscedasticity, or homogeneity of variances of the residuals, is one of the main assumptions of the regression analysis. We used White's general test for heteroscedasticity, which regresses the squared residuals on all distinct regressors, cross-products, and squares of regressors. It tests the null hypothesis that the variance of the residuals is homogenous. Low values of the derived Lagrange multiplier statistic and high values of p indicate that we

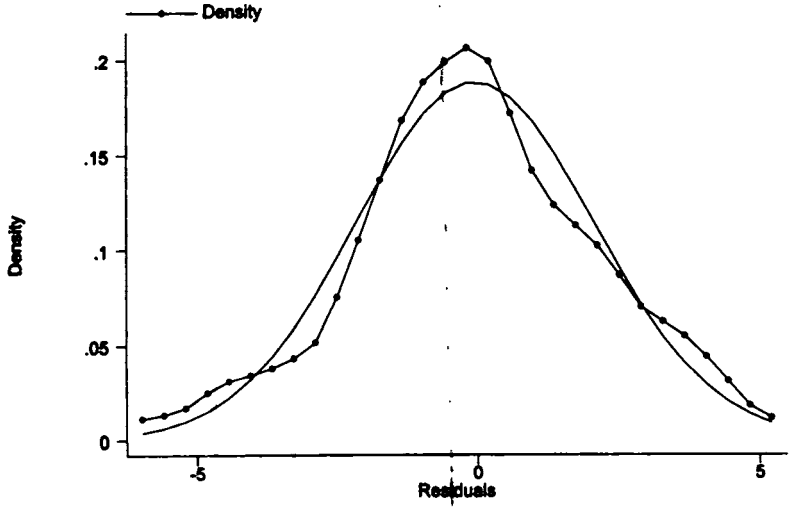


Figure 3.3. Kernel Density Estimate, which compares the estimated curve to the normal density (data for the Verbal Fluency–FAS test were used).

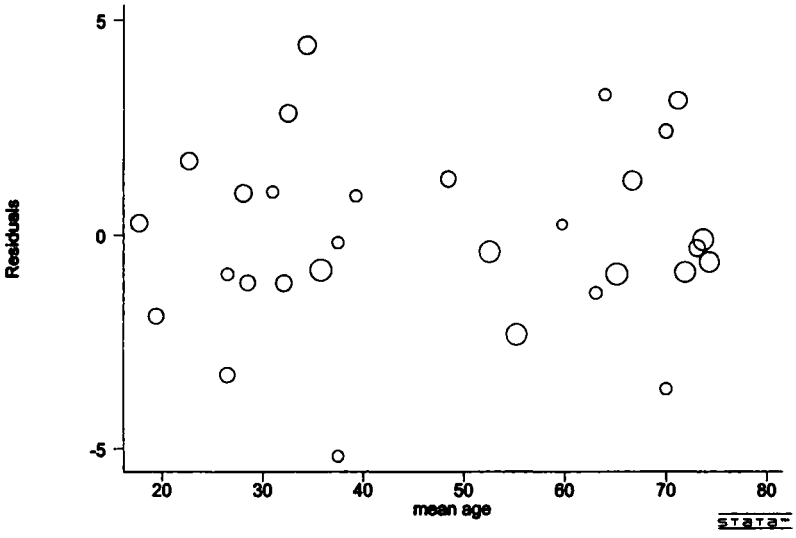


Figure 3.4. Plot of residuals regressed on age (data for the Verbal Fluency–FAS test were used). The size of the bubbles reflects the size of the standard errors of the data points, reciprocal to their weights.

cannot reject the hypothesis of homogeneity of variance in the residuals. A dispersion of residuals plotted vs. fitted values on the residual-vs.-fitted plot (rvf plot) was visually inspected for each regression. In a model with a good fit and homogenous residual variances, this distribution should have no pattern. An rvf plot for the FAS is included (see Fig. 3.5) to illustrate this technique. However, these plots are not reproduced in the meta-analytic tables in the appendices.

Independence assumption refers to the expectation that errors associated with one observation are not correlated with errors associated with any other observation. Our data clearly do not meet this assumption because the data points derived from the same study (e.g., when the scores are stratified by

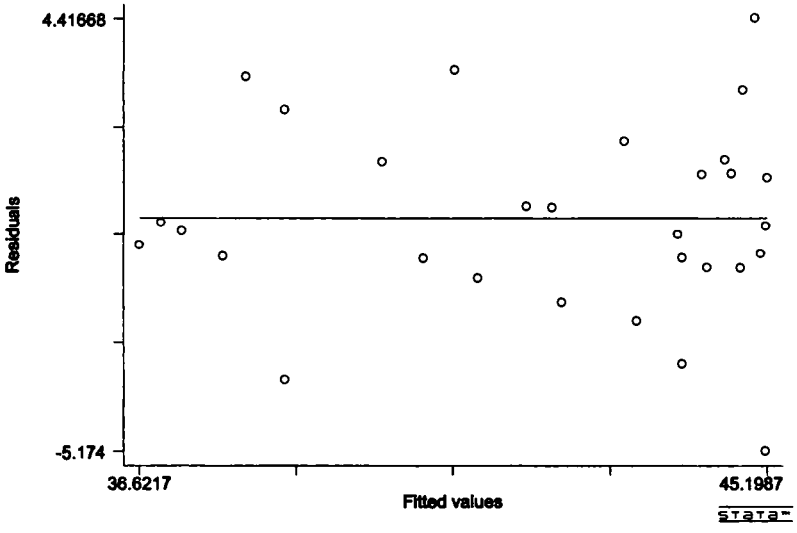


Figure 3.5. Residual-vs.-fitted plot (data for the Verbal Fluency–FAS test were used).

age group) are likely to be related and subjected to the same source of error. To account for the lack of independence, we used the cluster option for model estimation, which specifies that observations are independent across studies (clusters) but not within studies.

Effect of Demographic Variables

The effect of education was explored with the "metareg" command, which yields an estimated between-study variance tau^2, measuring residual heterogeneity adjusted for covariates. The value of the tau^2 estimate was compared for regressions of test means on additive components of variance with and without education. If the tau^2 value for the regression with education was much lower, indicating that education explains a considerable amount of heterogeneity in test performance, *education* was entered as a predictor variable into the equation used for the model estimation. If R^2 considerably improved as a result of addition of the *education* term and the t value for *education* was high with a low p value, the coefficient for *education*, derived from the latter regression, was used as a correction factor in the tables for relevant tests. Where education accounts for a large proportion of variance in test performance, the predicted scores listed in the age-stratified tables are accurate for individuals with education at the mean level for the original data set. With every year of education above or below the mean, expected gains or losses in test performance are equal to the coefficient for the *education* term. For example, values listed in the prediction table for the FAS are accurate for individuals with approximately 14 years of education since the mean education across all samples for this test is 14.31 (see Table 11m.1 in Appendix 11m, under "Description of the aggregate sample"). Thus, an expected score for a 37-year-old individual with 12 years of education (2 years below the mean of 14 cited above) is $45.17 - 2(0.50) = 44.17$.

Correction tables provided for the FAS and the Trailmaking Test (TMT) parts A and B allow such adjustments by adding or subtracting the appropriate correction factor to/from the predicted scores provided in the prediction tables. The SD to be used with the education-corrected score is that for the person's actual age group (which is relevant to the TMT tables but not to the FAS as the same SD is used for all age ranges for the latter). It should be noted that the range of years of education for the correction tables is limited. This limitation is due to a lack of empirical data for individuals with lower levels of education in the studies reviewed. We do not know whether the pattern of education/test performance relationship linearly continues into the lower educational levels. Therefore, extrapolation of the suggested correction pattern onto educational levels falling below the empirically supported range might undermine the accuracy of the prediction.

The effect of gender on test performance was assessed by adding a variable accounting for a percent of males in the sample as a predictor variable into a regression of test means on age. In addition, a t-test was run on the data that are reported for males and females separately. Male/female differences in mean test scores are reported in the tables, where appropriate. If a sufficient number of studies for a specific neuropsychological test report the data stratified by gender and a significant relationship between gender and test scores is highlighted in the literature, age-predicted scores are presented for males and females separately (e.g., GPT). For a number of neuropsychological tests, the differences between genders were not large enough to warrant separate predicted tables or addition of a correction factor for gender.

Although it is widely known that intelligence level makes a considerable contribution to performance on certain tests, we could not provide corrections for IQ level because of the paucity of reported data on IQ in the samples aggregated for the analyses in this book.

Similarly, the volume of information on ethnicity or other demographic variables gathered from the studies reviewed was not sufficient to conduct statistical analyses.

Comments on the Applicability of the Meta-Analyses Presented in This Book

As discussed earlier, an advantage of any meta-analysis is in increased power to direct informed clinical decisions based on synthesis of empirical data. Data derived from individual studies underlying our meta-analyses might be biased by imperfect sampling procedures, random individual differences due to small sample sizes in each demographic cell, and deviations from standard administration procedures. In addition, they are setting-specific and contain data for limited age ranges or demographic categories. Thus, choosing a normative data set as a reference for a specific patient might become a time-consuming undertaking. Meta-analytically derived regression estimates are based on large aggregate samples and represent the mathematical center of all studies across demographic groups. As such, regression-based tables of normative data are relatively free of chance factors affecting individual studies. However, regression-based norms should not be used as a substitute for empirically derived tables presented in the context of study reviews.

Any averaging results in a loss of specific qualities. We intended to present corrections for variables that are in theory expected to affect test performance. However, we were limited by the data available in the literature, which in many cases seem to be at odds with the theory. Individual data sets based on a sample of participants who are similar in terms of setting, demographic characteristics, and/or functional level to the patient for whom normative comparisons are sought would provide more accurate estimates of expected performance than regression-based tables. It has been emphasized by a number of investigators (e.g., Heaton et al., 1986; Kalechstein et al., 1998; Ross & Lichtenberg, 1998; Van Gorp & McMullen, 1997) that a selection of the normative data set should be guided by the comparability of the patient's demographic characteristics to those of the normative data and, more specifically, by the moderating variable that is most likely to affect performance (e.g., age for tests tapping psychomotor speed, education for tests emphasizing verbal achievement). Regression-based predictions are best considered an aid in selecting an appropriate table when results from different studies yield contradictory values.

Regression-based norms have been criticized (Fastenau, 1998; Fastenau & Adams, 1996; Heaton et al., 1996; Morgan & Caccappolo-van Vliet, 2001; Moses et al., 1999). Major criticisms refer to the concerns of violation of assumptions undermining the accuracy of prediction and extrapolation of the rule summarizing the relationship between predictor and outcome variables to the ranges of the predictor variable that are not supported by available data. As it follows from the above description of the procedures for heterogeneity and parameter specification testing applied in our analyses, the issue of violation of assumptions was closely attended to. In addition, all predicted values fall strictly within the empirically supported ranges of the predictor variables.

In spite of these efforts, the scope and quality of our analyses are limited by the scope of the data available in the literature. The accuracy of regression solutions presented in this book is undermined by several factors:

1. Age groupings provided in the literature vary greatly between studies. Whereas for one study the mean age of 48 years might represent a range of 45–50, in another study the mean age of 48 represents a range of 20–86. The performance score reported in the latter study is much less meaningful in terms of age-referenced prediction than in the former. This situation was mitigated by weighting data points on SEs for the means as this weighting takes into account the dispersion around the mean.
2. Evaluation of the effects of demographic variables on estimates of test performance was limited by scarcity of demographic data provided in the literature. For example, an important variable such as IQ, which is expected to contribute significantly to variability in several neuropsychological tests, had very

limited variance across the data sets. Only few studies reported IQ.

3. Levels of education and IQ for the majority of data sets are high. Therefore, the predicted values overestimate expected performance for individuals with a high school education or below and with average or lower than average range of intelligence.
4. We cannot describe our aggregate sample in terms of ethnic distribution because of scarcity of information on participants' ethnicity in the individual articles. We believe that the underlying samples are not representative of the mixture of ethnic groups according to U.S. Census figures since many samples were dominated by Caucasian participants. Those data that were collected exclusively on representatives of specific ethnic groups (e.g., Chinese, African American, or Hispanic) were not included in the meta-analyses as they increase the heterogeneity of the data. Ideally, separate analyses on data for different ethnic groups should be conducted in the future, providing that a sufficient number of studies reporting normative data specifically for different ethnic groups will be generated.
5. Increments in the values of predictor or moderator variables extracted from the literature are uneven. As reflected in scatterplots depicting the distribution of data points around the regression line for each relevant neuropsychological test, available data seem to cluster at the young and advanced ages, with more scarce data points in-between. Further investigations are needed to assure consistency in the relationship between predictor and outcome variables across all ages. However, large gaps in the ranges of predictor or moderator variables were avoided by eliminating extreme scores from the analyses. As a consequence of such adherence to empirically supported data, ranges of demographic categories covered in prediction tables are restricted; e.g., age groups are limited from both ends, and lower levels of education are not represented.
6. The suggested predictions for age (and education in a few cases) are based on the data for largely intact samples. It is unknown if the same relationship between demographic variables and test performance holds for individuals with brain pathology. Ultimately, normative databases should be expanded to include meta-analyses based on various clinical samples across test batteries, to acquire information on expected performance profiles for different diagnostic categories.

In spite of the weaknesses addressed above, we hope that the predictions presented in this book will facilitate the process of clinical decision making, which encompasses historical, clinical, and psychometric information.